Kee, K. F. (2017) The ten adoption drivers of open source software that enables e-research in data factories for open innovations. In S. A. Matei, N. Jullien, & S. P. Goggins (Eds.), *Big data factories, computational sciences: Collaborative approaches* (pp. 80-102). New York, Springer. https://link.springer.com/chapter/10.1007/978-3-319-59186-5_5

5. The Ten Adoption Drivers of Open Source Software That Enables e-Research in Data Factories for Open Innovations

Kerk F. Kee[1]

(1) School of Communication, Chapman University, Orange, CA 92866, USA

Kerk F. Kee
Email: kerk.kee@gmail.com

Keywords Open innovation – Adoption of innovation – Big data – Data dissemination – E-research – Scientific digital platforms

Introduction

According to the Oxford dictionary online, a factory is "[a] building or group of buildings where goods are manufactured or assembled chiefly by machine." To use the word "factory" in conjunction with "data," one can interpret the idea of "data factory" as a virtual arrangement or group of arrangements where big data sets are produced, aggregated, recombined, and/or repurposed mainly by cyberinfrastructure. The meta-platform of cyberinfrastructure includes open source software, visualization systems, remote instruments, distributed sensors, high-speed networks, supercomputers, communication technologies, and the multidisciplinary experts involved in the aggregation of big data and the production of knowledge based on the data (Atkin et al. 2003; Kee et al. 2011). Towns et al. (2014) refer to these also as advanced digital services for research and education with big data.

In the metaphor of a factory, an important point is that raw materials get turned into useful products through material manipulations and industrial treatments. Similarly, in a data factory, raw digital data get turned into meaningful insights through computational processing and data analysis. A critical component in data factory is the software that preprocesses and analyzes raw digital data. In fact, many results from the analysis of big data depend on and/or are tied to specific software applications. Therefore, insights drawn from big data are software dependent; without good and appropriate software, the hidden insights in big data cannot be fully tapped.

Additionally, the metaphor of a factory also conjures up the notion of "standardization," a practice made commonplace during the industrial revolution with the introduction of Taylor's scientific management and time and motion studies (Miller 2008). Standardization is important to the idea of data factories, as the standardization of data format and the interoperability of data make data aggregation, recombination, and repurposing for even larger-scale analysis possible. Therefore, a piece of good software should be designed to be easily adopted and widely diffused in order to facilitate the standardization and interoperability of data for data factories.

While much attention has been given to big data as the raw materials that have hidden insights, limited attention has been given to the open source software that turn raw materials into powerful insights. However, without

successful design, development, adoption, and implementation of useful software, raw materials will remain raw materials with hidden insights. Metaphorically, a factory full of raw materials without machines to process them is just that – a factory full of raw materials. A factory full of raw materials processed and assembled in a meaningful way can turn the rawness into usefulness. Therefore, intentional and strategic efforts should be carried out to promote wider adoption of good software applications.

The purpose of this chapter is to explore what drives the adoption and diffusion of open source software that can usher in the vision of data factories. With the adoption of good software applications across the community, researchers can begin moving individual data sets developed by independent projects across geographic locations and disciplinary domains into a broader data ecosystem sustainable over the long term. The data ecosystem should also be easily accessible and used by present and future researchers not directly involved with data collection and documentation of the individual data sets.

In order to achieve the stated goal for this chapter, it is organized with the following sections. First, the concepts of data, big data, and e-research are defined. Second, the largest National Science Foundation's (NSF) supercomputing consortium, XSEDE (Extreme Science and Engineering Discovery Environment), is discussed as a specific case of a data factory. Third, based on interviews conducted with community stakeholders of XSEDE, ten drivers of open source software adoption are discussed along with associated critical questions to promote intentional design of software for successful diffusion in the larger research community. Finally, a conclusion with implications wraps up the chapter.

Data, Big Data, and e-Research

Schroeder (2014) defines *data* as the materials that belong to the object(s) or phenomenon(a) of investigation and that data are the most useful unit of analysis for the investigation, which involves data collection before the interpretation. To take it further, Meyer and Schroeder (2015) argue that when a data set is a magnitude larger than any other existing data sets in size and scope within a given domain, the data set is qualified as *big data*. Furthermore, they suggest that big data represents a new form of collaborative interaction with and around materials for research. The idea is that big data do not exist simply as materials; they require multidisciplinary experts to collaborate in order to harness big data for important insights.

Besides the scholarly definition of big data offered by Schroeder and Meyer, big data is more commonly defined in the industries by several keywords that begin with the letter V. More specifically, the concept of big data was defined by what was first known as the three Vs of big data: volume, variety, and velocity (Laney 2001). The first V of *volume* refers to the size of the data, and it is often measured in terabyte and petabytes. This characteristic is almost intuitive, as the volume is what makes a data set big or bigger than other existing data sets in a given domain. The second V of *variety* indicates that big data have a range of data formats, often referred to as structured and/or unstructured data. If a big data set is made up of simply structured data, its aggregation, recombination, and analysis are relatively straightforward. If a big data set consists of mainly unstructured data, computational analysis will require a lot of data cleaning and conversion, in order to create format consistency (which is also known as the interoperability of data). This is critical for the need of recombining and repurposing of previously isolated data sets from independent projects. The third V of big data is *velocity*, which refers to the speed at which data are produced and processed. The production and processing of big data are usually in real time or near real time. It is also this characteristic that gives big data the currency and dynamic advantage over traditional dated and static data.

Recently, Gandomi and Haider (2015) further argue that big data possess three additional Vs of variability, veracity, and value. *Variability* describes the flow rates of big data as fluctuating, unpredictable, and erratic. The

fluctuation of big data's flow rates is due to the fact that big data sets usually are the aggregated results of data coming from various sources. Therefore, big data usually show periodic and sporadic ups and downs in flow rates. The next V of *veracity* implies that despite big data's inexactitude, imprecision, and uncertainty, they hold significant and hidden insights. The insights require strategic harnessing by humans and machines. Finally, the last V of *value* signifies that there is important worth that can be drawn from big data's large volume. As previously mentioned, the large volume of big data is the obvious defining characteristic of big data. Although the large volume of big data, often measured in terabytes and petabytes today, is commonly used as the primary definition of big data, Gandomi and Haider argue that the notion of volume is relative – what is regarded as big at the present time may be small in the future.

Given Gandomi and Haider's point above, perhaps the definition offered by Mayer-Schönberger and Cukier (2013) can be added to the list of defining characteristics. They argue that a data set is considered big data when the size of a sample drawn from a population is equal to the size of the entire population (i.e., N = all). Their argument stems from big data analytics' departure from the traditional practice of sampling and inferential statistics when it was impossible to obtain and/or analyze population data of an entire organization, community, country, or social system. Due to previous limitations in terms of data collection, researchers carefully drew a sample for analysis and then appropriately inferred from the sample certain insights about the population. This inference was determined by statistical calculations and probability. However, since population data can be obtained today, sometimes through passive data recording, there is no longer a need to simply draw a sample. Moreover, data analysis was previously limited to what a single computer can process. Given today's network capability, big data set can be processed by a network of supercomputers, such as in the case of the Extreme Science and Engineering Discovery Environment (XSEDE).

In summary, big data can be defined by volume, variety, velocity, variability, veracity, and value. These six Vs have also been reduced to simply the five Vs (volume, velocity, variety, value, and veracity) of big data. Today, the five Vs are widely used to define big data, such as in the call for papers by the 2016 IEEE Big Data conference in Washington DC. The main characteristic of volume can be understood also as when the size of the sample is equal to the size of the population or when the volume is at least

one magnitude bigger than the size and scope of other existing data sets within a given domain. Finally, big data present the need for multidisciplinary collaborations with and around the data.

What is the purpose of big data then? Meyer and Schroeder (2015) offer the answer that big data are being used for e-research (Borgman 2010; Dutton and Jeffreys 2010). They define *e-research* as "the use of shared and distributed digital software and data for the collaborative production of knowledge." They use the term e-research to be inclusive of e-science, computational social science, digital humanities, and any other computational analyses of big data for advancing knowledge by collaborative researchers. Interestingly, their definition of e-research has an emphasis on the collaborative nature of knowledge production. In other words, if a researcher simply digitalizes the data (e.g., scanning images of historical manuscripts for computational analysis) for personal use, and the researcher does not share the digitized manuscripts with a wider community of researchers, this researcher's work does not fully qualify as e-research. The emphasis of the collaborative nature of e-research is critical for the notion of data factories, as these factories are set up to support open innovations.

The challenge of volume can be addressed by high-performance computing (HPC) and/or high-throughput computing (HTC). When a data set is too big and a single desktop computer cannot process the data (i.e., choked and frozen when the "process" button is pushed), a researcher can apply for an allocation to access HPC and/or HTC at national resources, such as XSEDE. Therefore, the major challenge addressed by the data factory metaphor is that of variety. Metaphorically, a producer of goods made the goods from start to finish before the industrial revolution. Because the process was done by a holistic approach, each product was unique. While the uniqueness may be celebrated by some, the variety can be a problem when there is a need to aggregate, recombine, and repurpose them.

Taking a pro-innovation and innovation diffusion stance, this chapter presents the purpose of data factories as threefold. First, it is about standardization and interoperability to reduce the challenges that come with big data's variety and variability. Second, it is about having centralized data repositories and computational resources to process big data, supporting big data's volume and velocity. Finally, it is about creating and maintaining a thriving and collaborative community around open innovations, so big data's veracity and value can be fully realized. Given the purpose discussed, the

next section presents XSEDE as a specific case of data factory.

XSEDE as a Data Factory

The Extreme Science and Engineering Discovery Environment (XSEDE, www.xsede.org) is the largest supercomputing consortium that provides computational resources and expertise for data-intensive research and education in science, engineering, social sciences, and humanities in the USA. XSEDE consists of more than 20 supercomputers and resources for advanced visualization and analysis of big data. The consortium is led by the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, and it includes partner centers such as the Texas Advanced Computing Center at the University of Texas at Austin, the San Diego Supercomputer Center at the University of California at San Diego, and universities such as Purdue University and University of Southern California, to name a few (for a full list, please visit https://www.xsede.org/leaders). These partner institutions each contribute one or more allocatable services to the consortium.

XSEDE is funded by the Office of Advanced Cyberinfrastructure (OAC) of the NSF's Computer and Information Science and Engineering (CISE) Directorate to continue advancing NSF's efforts in providing a national infrastructure to support the e-research community and cyberinfrastructure ecosystem started by the TeraGrid (2001–2006) and TeraGrid2 (2006–2011) projects (for information on TeraGrid and TeraGrid2, please see Lawrence and Zimmerman 2007, Towns 2011, and Zimmerman and Finholt 2006). Launched in July 2011 and funded at about $125 million for 5 years, XSEDE transitioned into XSEDE2 in September 2016 for another 5 years with a new round of funding at about $92 million. Similar to TeraGrid and TeraGrid2, XSEDE provides all the resources and support at no cost to the e-research community. XSEDE2 will continue in this approach to support big data and open innovations in the USA.

The goal of XSEDE is not simply to provide supercomputing power; the goal also includes the goal to provide a comprehensive and cohesive set of distributed infrastructure, digital services, support services, and technical expertise to enable e-research and cyberlearning (Towns et al. 2014). Broadly, XSEDE has supported researchers in computational finance, genomics, epidemiology, digital humanities, and social network analysis. Notable examples of groundbreaking research supported by XSEDE include

a study of high-frequency trading in the US stock market (see O'Hara et al. 2014) and the hydrogen sorption in a metal organic framework (see Pham et al. 2013), to name a couple. More importantly, based on theoretical assumptions from classical and quantum mechanics, the use of XSEDE's predecessor was utilized for performing the simulation and prediction of the behavior of biomolecules, a study that led to a discovery that awarded Martin Karplus, Michael Levitt, and Arieh Warshel the 2013 Nobel Prize in Chemistry. Their computational simulation was innovative in taking chemistry research outside of the traditional laboratory (Towns et al. 2014).

XSEDE stores tens to hundreds of petabytes of data, supports a few hundreds of software packages, as well as provides training and services to more than 10,000 researchers and 2500 projects across all 50 states to harness big data for research discovery and knowledge production. XSEDE also supports international researchers from over 100 universities in more than 35 countries who collaborate with the US researchers XSEDE directly supports. XSEDE is an exemplar of a data factory, as Towns et al. (2014) explain – the purpose of XSEDE is for:

> Making codes run faster and more easily allows researchers to get more science done in a fixed amount of time. Lowering the barrier for access to and use of digital services enables additional research in established communities and in new communities who haven't harnessed these services to date. Such productivity increases can be the difference between an infeasible project and a feasible one, reducing the time to publishing scientific findings.

The notions of efficiency and productivity, two characteristics of the machine metaphor of industrial revolution (Miller 2008), are prominent in XSEDE.

As previously stated, the idea of "data factory" can be interpreted as a virtual arrangement or group of arrangements where big data sets are produced and aggregated mainly by cyberinfrastructure. A key feature of cyberinfrastructure is the open source software applications necessary for processing big data. Understanding the adoption drivers that promote diffusion of these software applications is important because existing efforts should not be wasted and new users do not need to reinvent the wheel. Furthermore, wider diffusion of good software will also help create

standardization and interoperability of data, further promoting the vision of data factories. Standardization can reduce idiosyncratic measures, and data formats, instead, move data from isolated projects and locked box repositories more easily into a longitudinal data warehouse associated with certain data factories. Finally, with wider adoption, more data can be aggregated, recombined, and integrated to perform analysis at unprecedented scale, to tackle big problems previously limited by the volume and variety of data and the limitation of existing software and supercomputing resources. The ultimate outcome of a pro-innovation diffusion effort in this sense can lead to more innovations and breakthroughs that benefit societies and humanity worldwide. In order to promote diffusion, the next section explores the ten drivers that promote the adoption of open source software for data factories and open innovations within the XSEDE community.

The Ten Adoption Drivers of Open Source Software in XSEDE

The ten drivers discussed in this section were identified in an analysis based on 135 in-depth interviews with domain researchers (as technology users), computational technologists (as software developers), and center administrators (as data center leaders) who consider themselves stakeholders of the XSEDE community (for more details, see Kee et al. 2016). The interviews were systematically analyzed using the grounded theory approach (Glaser and Strauss 1967; Kee and Thompson-Hayes 2012; Strauss and Corbin 1998). The ten drivers are also discussed with critical questions from the perspective of potential new users. These questions represent the kind of issues that stakeholders should keep in mind while designing and promoting their software within the larger research community to support data factories and open innovations.

Driven by Needs

The first adoption driver is the software's ability to meet users' existing needs. While research to date is still inconclusive about if users' needs drive innovation (see von Hippel 2005 on how lead users created innovations to meet their own needs) or an innovation creates a market for an unknown need (see Daly 2011 on how iPod created a completely new market), the development and adoption of open source software for big data are usually driven by known needs in the research community. This is because big data usually exist before the software to process them is available, and the software is designed to harness existing data. The segment that makes up the potential user market are busy professionals who do not have time to adopt a piece of software simply for personal enjoyment, but for a compelling reason, such as a pressing problem that represents a dire need for a solution.

Furthermore, the design and development of open source software can be very time consuming and financially expensive. This is why many software applications are developed by federally funded projects for 3–5 years (Kee and Browning 2010), such as those supported by NSF's OAC. In these projects, if the inception teams are not able to articulate a compelling rationale with clear reasons for the need to develop a piece of software for

research, the projects would not be funded by NSF and other federal agencies (such as the Department of Energy, National Institutes of Health, National Oceanic and Atmospheric Administration). The rationales are often based on grand challenges and critical problems well-documented in the research literature. Therefore, in order for a piece of open source software to widely diffuse, it needs to clearly meet the needs of potential users and the community/funders behind their work. In fact, in their discussion of XSEDE, Towns et al. (2014) open the article by stating that the establishment of XSEDE itself was "[d]riven by community needs." Therefore, a critical question stakeholders should keep in mind that a potential user may ask is "Does this software meet my needs?"

Organized Access

Once there is a compelling need, potential users require organized access to find the open source software they may adopt. The notion of organized access is not simply having an online link to download a piece of software; the notion includes having a systematically designed location (usually a website, such as HUBzero at https://hubzero.org/ and Galaxy Tool Shed at https://toolshed.g2.bx.psu.edu/) where inception teams post their software, active users rate, review, and comment on the software and potential adopters read about the software online easily. The website should be designed to facilitate a vibrant community where the interactions among different groups of stakeholders (inception teams, active users, and potential adopters) come together to carry a piece of open source software forward.

Having organized access to an online marketplace where the marketplace is well known is important for diffusion. This driver is important for data factories as the community of users need to participate in the marketplace in order to generate open innovations collectively. A piece of diffusing software has to have a strong web presence, and it can be located at a known marketplace that is open and organized for a community of users. Therefore, the critical questions stakeholders should keep in mind that a potential user may ask are "Is the software easily available?" and "Can I find the software at a known location?"

Trialability

The third adoption driver of a piece of diffusible open source software is that

it allows potential adopter to try it out before full adoption. Many open source software applications in e-research to date have a high degree of trialability because they are open sourced. These software applications are different from their propriety counterparts in that all the source codes are freely open, so interested developers and savvy users can add to the software and extend the software features based on their existing needs. In other words, being open sourced allows for open innovations and ecologically driven evolution of software. This is an important point for trialability because it is often during the open trials that potential adopters cultivate an understanding of the software and how it works, what it means for them in their particular contexts.

The notion of trialability is important for data factories because the notion of open innovation is eventually driven by open free trials and organic contributions. Within the community of e-research and open innovations, members subscribe to the open sharing philosophy. Once a piece of software is aligned with the potential users' philosophical orientation, the software should also be easily implemented for a trial without too much learning time. A steep learning curve will discourage adoption. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Can I try this software without much time investment?"

Well-Documented

Documentation refers to having a complete record of how the software was developed, instructions on integrating and using the software, the decisions that went into the design, the updates, and exemplars of how the software has been successfully used to solve different big data problems. A piece of software that is well documented not only offers potential adopters simply basic information to download the software, it offers a learning environment that is akin to a fully developed course on a piece of software. In other words, the documentation cannot be outdated and/or skeletal. Otherwise, another software with better documentation will likely attract more active users and potential adopters.

Being well documented is also an important characteristic for data factories, because the community members for open innovations are diverse, and the vision is to maintain long-term data and technologies that allow longitudinal analyses. Even when the pioneering stakeholders are no longer alive 100 years from now, their well-documented software applications can

continue being updated and used by future researchers. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Is the software well documented with a complete track record and robust user guides?"

Community Driven

Building on the adoption drivers of trialability and being well documented as discussed above, the more people can try out a piece of software with helpful documentations, the more active a community will develop around the software. The open sourced nature continues to manifest in the adoption driver of being community driven. The open source philosophy does not only give innovations freely to a marketplace, it empowers a community of stakeholders to rally around the software. The source codes are open online; this allows many savvy users, potential adopters, and interested developers to participate in trying out the software, integrating the software, fixing the bugs, updating the codes, improving its functionality, and extending its usage to new problems and contexts previously not considered by the inception team.

Shirky (2009) beautifully elaborates on Eric Raymond's notion of "a plausible promise" – the promise that the original developer will not take advantage of community contribution for personal financial gain. A plausible promise is what gives community members the reason to join and contribute to the community. The driver of community driven is fundamental to data factories for open innovations. It is also this driver that gives future adopters the confidence that the software will continue to thrive with the support of the community. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Is there a thriving community that will carry this piece of software forward for the long term?"

Observability

The next adoption driver is observability. The notion of observability manifests in terms of how often near peers talk about a piece of software (i.e., word of mouth), how frequently the software is showcased in research presentations and/or demonstrations at a conference (i.e., community visibility), and its success in enabling good research and producing useful results (i.e., citation index). The notion of observability based on the three

dimensions of word of mouth, community visibility, and citation index allows a piece of software to create the impression that the software has a strong potential to be useful for potential adopters.

The driver of observability is also important for data factories and open innovations because the contribution to and access to repositories depends on whether community members are aware of the software and related data archives. The more observable the software is, the more likely it will attract a group of stakeholders around it. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "What software are my peers using, and how are they using it?"

Relative Advantage

The adoption driver of relative advantage refers to a piece of software's ability to outperform an existing software in multifold. It is important to note that potential adopters are often entrenched in their existing technologies. Therefore, it is difficult or painful for them to transition. The new open source software has to offer a multifold advantage for potential adopters to overcome their resistance to avoid pain during a software transition.

In the case of open source software, a large segment of potential adopters are not existing users of other open source software, but potential adopters of the computational approach to gain insights from big data. In other words, these individuals have to be convinced not simply that the software is going to help them do their work better, but that the computational approach and big data will help them solve problems that are bigger in scale and more complex in scope or to solve a problem that otherwise cannot be solved with their existing technologies and approaches based on sampling techniques and samples drawn from larger populations of interest.

The driver of relative advantage is also important for data factories because the idea of open sharing an open innovation is still relatively new for the traditional research community grounded in individual credits for hiring, tenure, and promotion. The bundle of software, big data, and computational approach need to appear a lot more beneficial than the traditional way of doing research. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Is this software a lot better than what I have right now?"

Simplicity

Simplicity is key to successful software adoption. Very few people will take the time and effort to adopt a piece of complex software that is difficult to learn. There are always some die-heart users who believe that to fully do computational data processing, one needs to know the nitty gritty of programming and supercomputers. However, these individuals make up a small segment of the market place, possibly only those who are referred to as "innovators" (2.5% of total population) in Rogers' (2003) original diffusion model.

Instead of the need to learn how to program like those previously referred to as active and savvy users, there is now a steady effort in creating science gateways to lower the barrier of entry (Wilkins-Diehr et al. 2008). Science gateways are essentially open source software designed with a user-friendly interface. According to the XSEDE website:

> A Science Gateway is a community-developed set of software, applications, and data that are integrated via a portal or a suite of applications, usually in a graphical user interface, that is further customized to meet the needs of a specific community. Gateways enable entire communities of users associated with a common discipline to use national resources through a common interface that is configured for optimal use. Researchers can focus on their scientific goals and less on assembling the cyberinfrastructure they require. Gateways can also foster collaborations and the exchange of ideas among researchers.

As described above, with a science gateway, users can simply use the point-and-click method to navigate and use the software to process big data. According to Towns et al. (2014), more than 40% of XSEDE users in 2013 were users of one of more than 35 science gateways associated with XSEDE in the same year. This portion of users is expected to continue growing over time. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Is this software simple to learn and easy to use?"

Compatibility

The adoption driver of compatibility refers to a piece of software's fit with a potential adopter's technological repertoire, behavioral practices, and

ideological orientation toward data-driven research. If the innovation is disruptive (technologically, behaviorally, and ideologically), both for the potential adopters and/or their collaborators, the innovation will suffer greatly in terms of compatibility. As today's researchers are heavily dependent on their technologies, the further a new piece of software departs from their existing routine and/or the norms in their disciplines, the more difficult it is for the software to be adopted.

This driver is also important for data factories because in order for a community of open innovations to thrive, it needs to attract many members. A potential member may compare and contrast if his/her data format is compatible with the format chosen by a data factory of interests. Without data interoperability, the aggregation of data sets into a big data set is difficult. The software and the data format go hand in hand for the adoption decision by potential users. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Can I easily integrate this software into my existing routine and collaborations?"

Adaptability

Traditionally, adoption with a deviation from the original purpose of a piece of software is considered as "noise" in diffusion research. This bias is understandable because a deviation does not count as a full adoption if a researcher or manufacturer is interested in tracking "successful adoption" of a new technology as originally designed. However, in the Web 2.0 era, a deviation from the original purpose (such as in terms of adaptability, repurposing, and reinvention) may aid in a piece of open source software's ability to diffuse. In other words, a piece of software's ability to adapt and be repurposed for a new problem and/or a new context may promote its wider adoption ultimately.

The adoption driver of adaptability should not be left as simply a happy accident. In fact, it can be an intentional diffusion strategy – a piece of software is designed to repurpose across problems, contexts, fields, and domains. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Can I take this piece of software from that domain and bring it into my domain?" Table 5.1 below summarizes the ten adoption drivers and associated critical questions as discussed above.

Table 5.1 The ten adoption drivers of open source software in the e-research community for data

factories and open innovations

Adoption drivers	**Critical questions**
Driven by needs	*“Does this software meet my needs?”*
Organized access	*“Is this software easily available?” and “Can I find the software at a known location?”*
Trialability	*“Can I try this software without much time investment?”*
Well documented	*“Is the software well documented with a complete track record and robust user guides?”*
Community driven	*“Is there a thriving community that will carry this piece of software forward for the long term?”*
Observability	*“What software are my peers using, and how are they using it?”*
Relative advantage	*“Is this software a lot better than what I have right now?”*
Simplicity	*“Is this software simple to learn and easy to use?”*
Compatibility	*“Can I easily integrate this software into my existing routine and collaborations?””*
Adaptability	*“Can I take this piece of software from that domain and bring it into my domain?”*

Conclusion, Discussion, and Implications

This chapter set out to explore the definitions of data, big data, and e-research in the context of data factories for open innovations. The metaphor of a factory for data is compelling as it implies key characteristics for data such as standardization and interoperability and for open innovations such as efficiency and productivity. The chapter presents the case of XSEDE as an exemplar of a data factory in the USA. Most importantly, this chapter laid out ten drivers that promote the adoption and diffusion of open source software in the e-research community to usher in the vision of data factories and open innovations. It is important to note that the ten drivers make up a need-based diffusion model, a broader technology adoption framework. Although the ten drivers were presented in a linear and sequential way, it is important to keep in mind that they are interconnected and they influence each other in a complex way at any given time.

The topic of software adoption is not simply a theoretical question; it is also an important practical question. Instead of providing direct recommendations for practice, the ten drivers were presented with associated critical questions (see Table 5.1 for a summary) to prompt the stakeholders to ponder upon and discussed the ten different drivers at any given point in time. In a fast-changing world of technologies, a specific recommendation is likely to be outdated in the foreseeable future. Furthermore, a recommendation that works well in one particular disciplinary domain may not work in another domain. However, by engaging with the critical questions, stakeholders can come up with the best answers for themselves in their given contextual and historical contexts. Therefore, the critical questions are useful for facilitating stakeholders' regular reflections on the challenges and opportunities to promote their software applications for data factories and open innovations.

While the focal point was on the adoption of open source software as a technology, an important insight stemmed from the discussion above is that the adoption decisions are multidimensional. Kee (2017) uses the adoption of green technologies within the workplace as an example to make this point. More specifically, the adoption of the green technologies also involves the adoption of the recycling and/or conservation behaviors and the belief and mindset that environmental sustainability is of critical urgency and

importance within the workplace. If the push to adopt a green technology only focuses on the technology itself, the stakeholders are missing the critical fact that the adoption of the technology is not complete without the adoption of the associated behavioral practices and philosophical ideologies.

Similarly, the argument can be extended to the adoption of open source software for data factories and open innovations. The potential adopters need to be willing to modify existing practices to make the software fit into existing routines and collaborations. The potential adopters also need to strongly believe that open source software, data factories, and open innovations are the ways of the future of research and knowledge production. The adoption decision is multidimensional, as it involves the adoption of the material objects (i.e., open source software, big data), the behavioral practices (i.e., large-scale scientific collaborations, open sharing of data and documentation), and philosophical ideologies (i.e., data factories, open innovations). The adoption of one dimension without the others would be considered incomplete. The case of XSEDE presents an interesting context to study the diffusion of multidimensional innovations for adoption.

Acknowledgment

The author thanks Mona Sleiman, Rion Dooley, Nancy Wilkins-Diehr, and John Towns for their support of this project. This research was funded by NSF ACI 1322305.

References

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., et al. (2003). Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-ribbon Advisory Panel on Cyberinfrastructure. Washington, DC: National Science Foundation. Retrieved from 19 Dec 2006. http://www.communitytechnology.org/nsf_ci_report/.

Borgman, C. L. (2010). *Scholarship in the digital age: Information, infrastructure, and the internet*. Cambridge, MA: MIT press.

Daly, J. A. (2011). *Advocacy: Championing ideas and influencing others*. New Haven: Yale University Press.

Dutton, W. H., & Jeffreys, P. W. (2010). *Worldwide research: An introduction*. Cambridge, MA: MIT Press.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*, 137–144.
[Crossref]

Glaser, B. G., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Piscataway: Aldine Transaction.

Kee, K. F. (2017). Adoption and diffusion. In C. Scott & L. Lewis (Eds.), *International encyclopedia of organizational communication, 1* (pp. 41–54). Chichester: Wiley-Blackwell.

Kee, K. F., & Browning, L. D. (2010). The dialectical tensions in the funding infrastructure of cyberinfrastructure. *Computer Supported Cooperative Work, 19*, 283–308.
[Crossref]

Kee, K. F., Cradduck, L., Blodgett, B., & Olwan, R. (2011). Cyberinfrastructure inside out: Definitions and influencing forces shaping its emergence, development, and implementation. In D. Araya, Y. Breindl, & T. Houghton (Eds.), *Nexus: New intersections in Internet research* (pp. 157–189). New York: Peter Lang.

Kee, K. F., & Thompson-Hayes, M. (2012). Conducting effective interviews about virtual work: Gathering and analyzing data using a grounded theory approach. In S. D. Long (Ed.), *Virtual work and human interaction research* (pp. 192–212). Hershey: IGI Global.
[Crossref]

Kee, K. F., Sleiman, M., Williams, M., & Stewart, D. (2016). The 10 attributes that drive adoption and diffusion of computational tools in e-science. In P. Navrátil, M. Dahan, D. Hart, A. Romanella, & N. Sukhija (Eds.), *Proceedings of the 2016 XSEDE Conference: Diversity, big data, & science at scale*. New York: ACM.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*, 70.

Lawrence, K. A., Zimmerman, A. (2007). TeraGrid planning process report: August 2007 user workshops. *Collaboratory for Research on Electronic Work, School of Information, University of Michigan.* Retrieved January 10, 2010, from http://deepblue.lib.umich.edu/handle/2027.42/61842

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Meyer, E. T., & Schroeder, R. (2015). *Knowledge machines: Digital transformations of the sciences and humanities*. Cambridge, MA: MIT Press.

Miller, K. (2008). *Organizational communication: Approaches and processes* (5th ed.). Belmont: Wadsworth Publishing.

O'Hara, M., Yao, C., & Ye, M. (2014). What's not there: Odd lots and market data. *The Journal of Finance, 69*, 2199–2236. https://doi.org/10.1111/jofi.12185.
[Crossref]

Pham, T., Forrest, K. A., Nugent, P., Belmabkhout, Y., Luebke, R., Eddaoudi, M.,. .. Space, B. (2013). Understanding hydrogen sorption in a metal–organic framework with open-metal sites and amide

functional groups. *The Journal of Physical Chemistry C, 117*, 9340–9354. doi: https://doi.org/10.1021/jp402304a.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.

Schroeder, R. (2014). Big data and the brave new world of social media research. *Big Data & Society, 1*, 2053951714563194.
[Crossref]

Shirky, C. (2009). *Here comes everybody: The power of organizing without organizations*. New York: Penguim.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks: Sage.

Towns, J. (2011). *The sunset of TeraGrid and the dawn of XSEDE*. Paper presented at the The 10th Annual Meeting on High Performance Computing and Infrastructure in Norway (NOTUR2011), Oslo. http://www.notur.no/notur2011/material/TG-to-XSEDE-for-NOTUR11_Towns.pdf.

Towns, J., Cockerill, T., Dahan, M. F., Ian, Gaither, K., Grimshaw, A., Hazlewood, V., … Wilkins-Diehr, N. (2014). XSEDE: Accelerating scientific discovery. *Computing in Science & Engineering, 16*, 62–74.

von Hippel, E. (2005). *Democratizing innovation*. Cambridge, MA: The MIT Press.

Wilkins-Diehr, N., Gannon, D., Klimeck, G., Oster, S., & Pamidighantam, S. (2008). TeraGrid science gateways and their impact on science. *Computer, 41*(11), 32–41.
[Crossref]

Zimmerman, A., & Finholt, T. A. (2006). TeraGrid user workshop final report. *Collaboratory for Research on Electronic Work, School of Information, University of Michigan.* Retrieved January 10, 2010, from http://deepblue.lib.umich.edu/handle/2027.42/61841.

Sorin Adam Matei, Nicolas Jullien and Sean P. Goggins (eds.), Big Data Factories, Computational Social Sciences, https://doi.org/10.1007/978-3-319-59186-5_6